In [13]:

```
import pandas as pd
import math
import matplotlib.pyplot as plt
import seaborn as sns
```

In [ ]:

```

```

In [17]:

```
Y = Data.Expenditures # the dependent variable
X = Data.Age # the independent variable
```

In [18]:

```
#coefficient b
b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
print("Value of b is: ",b)
```

```
Value of b is:  -0.33359609660627854
```

In [19]:

```
X_bar = Data.Age.mean() # sample mean of age
Y_bar = Data.Expenditures.mean() # sample mean of expenditures print("Mean Ag
print("Mean Expenditure : ", Y_bar)
```

```
Mean Expenditure :  101.11538461538461
```

In [20]:

```
 a = Y_bar - b*X_bar
print("Value of a : ", a)
```

```
Value of a :  114.24110795493165
```

In [22]:

```
Data["error"] = Data.Expenditures - a - b*Data.Age
Data.head()
```

Out[22]:

	Observation	Age	Expenditures	Unnamed: 3	Unnamed: 4	error
0	1	49	95	NaN	NaN	-2.894899
1	2	15	104	NaN	NaN	-5.237167
2	3	43	91	NaN	NaN	-8.896476
3	4	45	98	NaN	NaN	-1.229284
4	5	40	94	NaN	NaN	-6.897264

In [23]:

```
sum_sq_error = (Data.error ** 2).sum() # calculating the sum of squares
```

In [24]:

```
## calclating ci in the dataset
Data["c"] = (Data.Age - X_bar) / ((Data.Age - X_bar)**2).sum()
Data.head(6) # showing the first few rows of the enhanced dataset
```

Out[24]:

	Observation	Age	Expenditures	Unnamed: 3	Unnamed: 4	error	c
0	1	49	95	NaN	NaN	-2.894899	0.003411
1	2	15	104	NaN	NaN	-5.237167	-0.008603
2	3	43	91	NaN	NaN	-8.896476	0.001291
3	4	45	98	NaN	NaN	-1.229284	0.001998
4	5	40	94	NaN	NaN	-6.897264	0.000231
5	6	35	107	NaN	NaN	4.434755	-0.001536

In [25]:

```
beta = b - (Data.c * Data.error).sum()
print("The value of beta is: ", beta)
```

```
The value of beta is:  -0.33359609660628065
```

In [27]:

```
n = Data.shape[0] # number of entries
s_b_sq = np.sqrt(((Data.error)**2).sum()) / ((n-2) * (((X - X_bar)**2).sum()
print("The value of standard error is: ", s_b_sq)
```

```
The value of standard error is:  0.09536918278863911
```

In [28]:

```
t_b = (b)/s_b_sq
print("The t value of b is: ", t_b)
```

```
The t value of b is:  -3.4979443762835545
```

In [29]:

```
#Answer 1 summary
print("Summary of Answer a results\n")
print("Value of a : ", a)
print("Value of b : ",b)
print("The standard error is: ", s_b_sq)
print("The t value of b is: ", t_b)
```

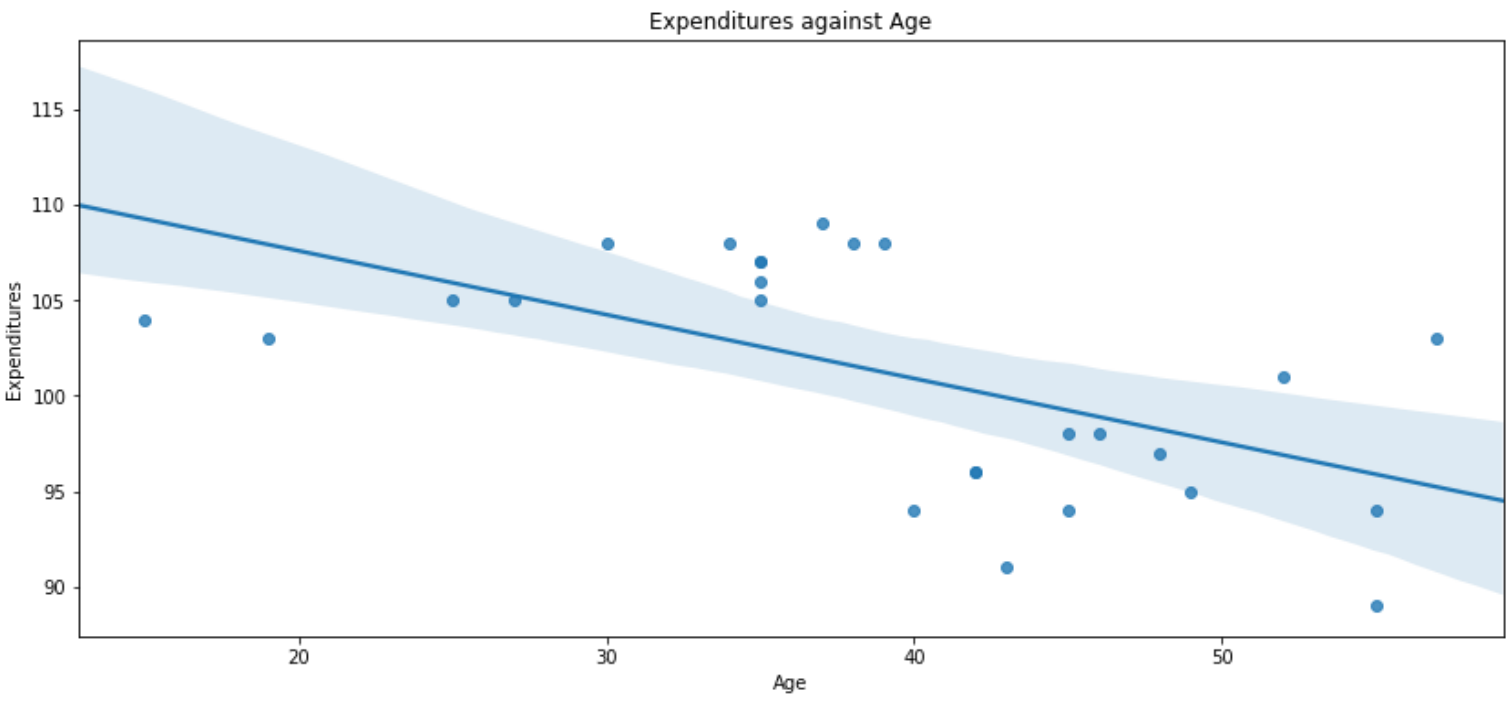
```
Summary of Answer a results

Value of a :  114.24110795493165
Value of b :  -0.33359609660627854
The standard error is:  0.09536918278863911
The t value of b is:  -3.4979443762835545
```

In [30]:

```
#Question b
plot = sns.regplot(data=Data, x= "Age", y= "Expenditures")
plot.figure.set_size_inches(14,6)
plot.axes.set_title('Expenditures against Age')
```

Out[30]:

```
Text(0.5, 1.0, 'Expenditures against Age')
```



Answer b

we can see that there is a decreasing trend between the age and expenditure. As the age decreases, expenditure also decrease. Additionally, we can also see two clusters of data around age groups less than 40 and greater than 40. These clusters will be analyzed further in next questions

In [31]:

```
lt40 = Data.Age < 40
Data_lt40 = Data[lt40].copy()
Data_lt40
```

Out[31]:

	Observation	Age	Expenditures	Unnamed: 3	Unnamed: 4	error	c
1	2	15	104	NaN	NaN	-5.237167	-0.008603
5	6	35	107	NaN	NaN	4.434755	-0.001536
7	8	38	108	NaN	NaN	6.435544	-0.000476
9	10	30	108	NaN	NaN	3.766775	-0.003303
13	14	25	105	NaN	NaN	-0.901206	-0.005070
14	15	35	107	NaN	NaN	4.434755	-0.001536
15	16	35	106	NaN	NaN	3.434755	-0.001536
16	17	35	105	NaN	NaN	2.434755	-0.001536
17	18	27	105	NaN	NaN	-0.234013	-0.004363
19	20	37	109	NaN	NaN	7.101948	-0.000829
21	22	19	103	NaN	NaN	-4.902782	-0.007190
24	25	34	108	NaN	NaN	5.101159	-0.001889
25	26	39	108	NaN	NaN	6.769140	-0.000122

In [34]:

```
## calculating the value of b which is needed to derive a
Y = Data_lt40.Expenditures # the dependent variable
X = Data_lt40.Age # the independent variable
b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
X_bar = X.mean() # sample mean of age
Y_bar = Y.mean()
a = Y_bar - b*X_bar
print("Summary of Answer c - part 1 results\n")
print("Value of a is: ", a)
print("Value of b is", b)
## calculate error from a and b
Data_lt40["error"] = Y - a - b*X
sum_sq_error = (Data_lt40.error ** 2).sum() # calculating the sum of squares
n = Data_lt40.shape[0] # number of entries
s_b_sq = np.sqrt(((Data_lt40.error)**2).sum()) / ((n-2) * (((X - X_bar)**2).
t_b = (b)/s_b_sq
print("The standard error is: ", s_b_sq)
print("The t value of b is: ", t_b)
# sample data set with errors and c
print("\n\n Sample data for the final dataset for Age less than 40 with error
Data_lt40.head()
```

```
Summary of Answer c - part 1 results

Value of a is:  100.23227718258495
Value of b is 0.1979712787782071
The standard error is:  0.044383667586451 25
The t value of b is:  4.460453350156233


 Sample data for the final dataset for Age less than 40 with error a
nd c
```

Out[34]:

	Observation	Age	Expenditures	Unnamed: 3	Unnamed: 4	error	c
1	2	15	104	NaN	NaN	0.798154	-0.008603
5	6	35	107	NaN	NaN	-0.161272	-0.001536
7	8	38	108	NaN	NaN	0.244814	-0.000476
9	10	30	108	NaN	NaN	1.828584	-0.003303
13	14	25	105	NaN	NaN	-0.181559	-0.005070

In [35]:

```
gt40 = Data.Age >= 40
Data_gt40 = Data[gt40].copy()
Data_gt40
```

Out[35]:

	Observation	Age	Expenditures	Unnamed: 3	Unnamed: 4	error	c
0	1	49	95	NaN	NaN	-2.894899	0.003411
2	3	43	91	NaN	NaN	-8.896476	0.001291
3	4	45	98	NaN	NaN	-1.229284	0.001998
4	5	40	94	NaN	NaN	-6.897264	0.000231
6	7	42	96	NaN	NaN	-4.230072	0.000938
8	9	46	98	NaN	NaN	-0.895688	0.002351
10	11	52	101	NaN	NaN	4.105889	0.004472
11	12	55	89	NaN	NaN	-6.893323	0.005532
12	13	42	96	NaN	NaN	-4.230072	0.000938
18	19	48	97	NaN	NaN	-1.228495	0.003058
20	21	45	94	NaN	NaN	-5.229284	0.001998
22	23	57	103	NaN	NaN	7.773870	0.006238
23	24	55	94	NaN	NaN	-1.893323	0.005532

In [37]:

```
## calculating the value of b which is needed to derive a
Y = Data_gt40.Expenditures # the dependent variable
X = Data_gt40.Age # the independent variable
b = ((X*Y).mean() - X.mean()*Y.mean()) / ((X**2).mean() - (X.mean())**2)
X_bar = X.mean() # sample mean of age
Y_bar = Y.mean()
a = Y_bar - b*X_bar
print("Summary of Answer c - part 2 results\n")
print("Value of a is: ", a)
print("Value of b is", b)
## calculate error from a and b
Data_gt40["error"] = Y - a - b*X
sum_sq_error = (Data_gt40.error ** 2).sum() # calculating the sum of squares
n = Data_gt40.shape[0] # number of entries
s_b_sq = np.sqrt(((Data_gt40.error)**2).sum()) / ((n-2) * (((X - X_bar)**2).
t_b = (b)/s_b_sq
print("The standard error is: ", s_b_sq)
print("The t value of b is: ", t_b)

# sample data set with errors and c
print("\n\n Sample data for the final dataset for Age greater than or equal t
Data_gt40.head()
```

```
Summary of Answer c - part 2 results

Value of a is:  88.87188902488657
Value of b is 0.14647082823339977
The standard error is:  0.19738441872591267
The t value of b is:  0.7420587155705977


 Sample data for the final dataset for Age greater than or equal to
40 with error and c
```

Out[37]:

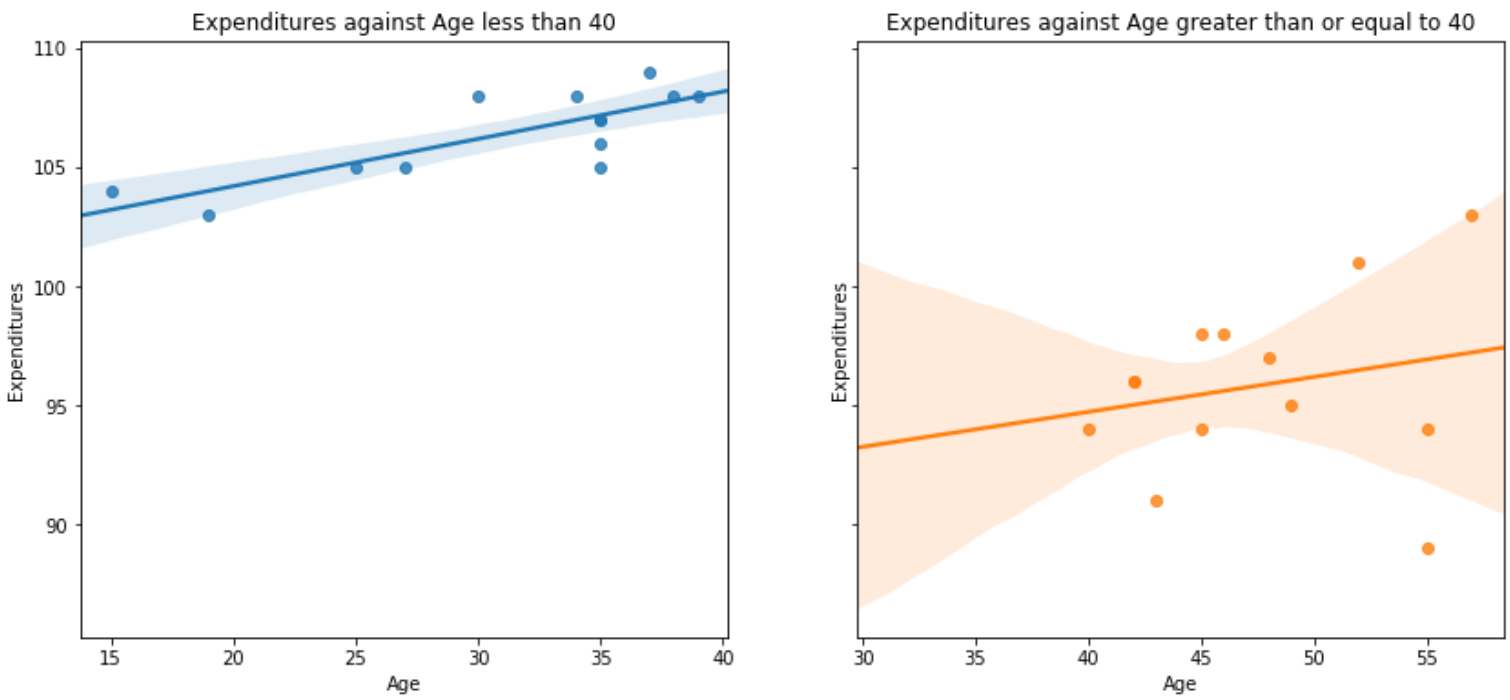
	Observation	Age	Expenditures	Unnamed: 3	Unnamed: 4	error	c
0	1	49	95	NaN	NaN	-1.048960	0.003411
2	3	43	91	NaN	NaN	-4.170135	0.001291
3	4	45	98	NaN	NaN	2.536924	0.001998
4	5	40	94	NaN	NaN	-0.730722	0.000231
6	7	42	96	NaN	NaN	0.976336	0.000938

In [38]:

```
fig, (ax1, ax2) = plt.subplots(ncols=2, sharey=True)
sns.regplot(x = Data_lt40.Age, y = Data_lt40.Expenditures, ax = ax1)
ax1.figure.set_size_inches(14,6)
ax1.axes.set_title('Expenditures against Age less than 40')
sns.regplot(x = Data_gt40.Age, y = Data_gt40.Expenditures, ax = ax2)
ax2.figure.set_size_inches(14,6)
ax2.axes.set_title('Expenditures against Age greater than or equal to 40')
```

Out[38]:

```
Text(0.5, 1.0, 'Expenditures against Age greater than or equal to 40
')
```



Answer d

Splitting the data into the two clusters mentioned in answer b gives opposite inference to what was seen in answer b. There is now an increasing relationship between age and expenditure i.e. as age increases, expenditure also increases. The reason for the overall decreasing trend was because the expenditure of people with age greater than 40 is generally lower than those with age less than 40 within the two clusters, people with age less than 40 have more sensitive spending habits. The curve for this cluster is steeper compared to the other which indicates that the effect of age on the spending habits is more in age group less than 40 compared to age group greater than 40

In [ ]:

```
1
```